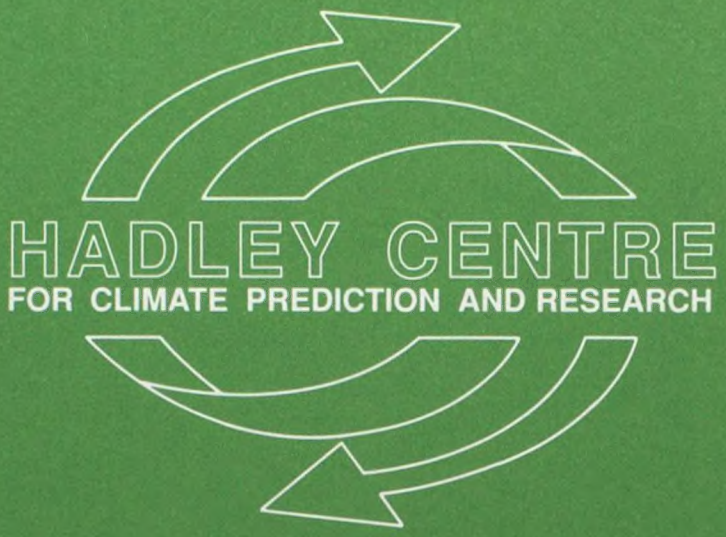

LEPS scores for assessing Climate model simulations and long-range forecasts.

by

C. K. Folland

CRTN 33

November 1992

CLIMATE
RESEARCH
TECHNICAL
NOTE

Hadley Centre
Meteorological Office
London Road
Bracknell
Berkshire RG12 2SY

CLIMATE RESEARCH TECHNICAL NOTE NO 33

LEPS SCORES FOR ASSESSING CLIMATE MODEL SIMULATIONS AND LONG-RANGE FORECASTS

by

Chris Folland

Nov 1992

Hadley Centre for Climate Prediction and Research
Meteorological Office
London Road
Bracknell
Berkshire RG12 2SY
U.K.

NOTE: This paper has not been published. Permission to quote it should be obtained from the Director of the Hadley Centre

Abstract

The experimental, equitable, Linear Error in Probability Space (LEPS) scoring system used in Ward and Folland (WF) (1991) is further developed after discussing some problems with the WF version of LEPS (here called LEPS1). A revised formulation, LEPS2, is derived that does not show the problem, exhibited by LEPS1, that scores for very poor forecasts can be slightly greater than those for slightly better forecasts. Such `bending back' of LEPS scores occurs when categorical forecasts are issued with a relatively high resolution in the categories or when the continuous form is used. Some properties of LEPS2 are compared with those of LEPS1 and a limited comparison is made with the equitable Folland-Painting scoring system. The latter is used for assessing monthly forecasts for the UK and also suffers from "bending back".

It can be desirable to adjust LEPS2 scores to provide a percentage skill score. Two kinds of percentage skill score are discussed, one of which penalises forecasts that are issued in category form by an amount that increases as the categories become wider. Although the equitable nature of the original LEPS2 scores is lost in percentage skill scores, such scores become "almost equitable" when a large enough ensemble of forecasts is assessed. Percentage skill scores may be especially useful for assessing simulations or predictions of continuous meteorological fields.

1. Introduction

LEPS arose from a need to assess the three main types of long-range forecasts on a common basis. These are: (a) best estimate forecasts that predict a single value of a variable from a continuous distribution, assessed using LEPSCONT; (b) best-estimate forecasts of one of a set of pre-defined categories, assessed using LEPSCAT and (c) forecasts of the probabilities of all pre-defined categories, assessed with LEPSPROB. LEPS measures skill relative to a reference "chance forecast" strategy, though other reference forecast strategies could be used. The assessments used for monthly Met Office long-range forecasts for the UK are the Folland-Painting (FP) scores (Folland et al., 1986). The FP scores are based on the negative logarithm of the distance between the forecast and observation, measured in the cumulative probability distribution of the observations. This quantity is normalised to create an equitable scoring system so that the chance score for all predefined forecast categories is zero and the score for a constant forecast is zero. Like LEPS1, the FP system is "doubly equitable" as the chance score for all observed categories is also zero giving a symmetrical scoring matrix. Thus the observed bunching in the occurrence of extreme or near normal categories does not, of itself, lead to fluctuations in forecast skill.

FP scores are based on information theory. LEPS scores, on the other hand, are based on the distance between the forecast and the observation. The FP system can only assess forecasts issued in discrete categories as it becomes undefined when continuous variables are forecast. Thus if the forecast and the observation occupy the same point in the cumulative probability distribution, their distance apart is zero and the non-normalised FP score is -log(0) and infinite. Normalised FP scores also "bend back" when octiles or smaller categories are used, giving scores for maximally incorrect forecasts that are slightly less negative than those for less erroneous forecasts. This problem occurs with the experimental LEPS1 scores, as mentioned by WF. Although no problem arose with their published quint scoring table, this is a "design fault".

Arguments can be marshalled against long-range forecasts of continuous variables; in fact, the FP system is based on the idea that, in the face of uncertainty, discrete categories should be forecast. However, it is felt to be useful to have a method of relating the "skill" of multiple regression forecasts of a single "best estimate" value, taken from a continuous distribution, to that of a forecast of a discrete category derived from the same value. A score that can assess forecasts of continuous variables also has potential for assessing climate model simulations or predictions of fields of continuous variables such as surface pressure. It may sometimes be more appropriate to measure the errors of such simulations according to their errors in the climatological probability distribution, as done by LEPS, than by using the root mean squared error (RMS) or correlation. LEPS penalises errors in the simulation of an excessively low surface pressure less than would RMS but gives a greater relative penalty to a small error in a near normal pressure simulation. Correlation has the fault that no account is taken of any systematic differences between the variance of the simulations and the observations. Its values are also very sensitive to small differences between forecasts and observations when both are near the climatological value, tending to give an artificially low skill. Potts (1991) gives a detailed discussion.

A comprehensive discussion of scoring systems is given by Daan (1984). He shows that scoring is a complex subject with many pitfalls. Equitable skill scores are discussed by Gandin and Murphy (1992). A discussion of features of LEPS and other scoring systems is contained in Potts (1991).

2. Derivation of LEPS1

The key idea is to measure the error in a forecast according to the distance between the position of the forecast and corresponding observation in units of their respective cumulative probability distributions. The cumulative probability of the forecasts and observations may not be the same and can have any shape consistent with a monotonic increase in value when plotted against increasing forecast or observation values. Thus the cumulative probability of monthly UK forecasts issued prior to about 1981 was radically different from that of the observations and was still somewhat different afterwards. This mismatch can markedly effect assessed skill and has not been allowed for up to now in assessments of UK monthly forecasts as the cumulative probability of the forecasts and observations was assumed to be the same. The FP system, as well as LEPS, can nevertheless allow for this by using a separate cumulative frequency distribution for the forecasts based on a recent period or even just for the period of assessment. Thus the position of the forecast is measured in its own cumulative probability distribution and compared with that of the observation in its different cumulative probability distribution. This procedure can probably only be justified, however, if a suitable test shows the distributions to be significantly different. Tables of quint or other category scores remain unchanged as only the value of the forecast variable corresponding to each category boundary alters.

In the basic, non equitable, form of LEPS, a perfect forecast is given a score a. For imperfect forecasts, a is reduced by a penalty equal to the modulus of the difference between the position, P_v, of the observation in the cumulative probability distribution and that of the forecast, P_f. P_v and P_f are fractions with a lowest possible value of zero and highest value of unity. Such a score has the form:

$$S_{fv} = a - f(|P_v - P_f|) \quad \ldots\ldots\ldots\ldots (1a)$$

In LEPS the function f is set to unity. The form used in LEPS1 sets a=1 giving:

$$S_{fv} = 1 - |P_v - P_f| \quad \ldots\ldots\ldots\ldots (1b)$$

S_{fv} needs to be normalised relative to chance so that the chance score is zero for any forecast or observed category to give an equitable score S'_{fv} using a standard method of normalising a contingency table. The normalising factor is the product of the averages of all possible scores for the given forecast category, and of all possible scores for the observation category, divided by the grand mean score:

$$S'_{fv} = S_{fv} - \frac{\overline{S}_f \overline{S}_v}{\overline{S}_{fv}} \quad \ldots\ldots\ldots\ldots (2a)$$

In the continuous case this is:

$$S'_{fv} = S_{fv} - \frac{\int S_v dP_f \int S_f dP_v}{\iint S_{fv} dP_f dP_v} \quad \ldots\ldots\ldots\ldots (2b)$$

Equation 2 is evaluated by imposing conditions to ensure that S is always positive:

(i) $S_{fv} = 1-(P_f-P_v)$ $\qquad P_f > P_v$
(ii) $S_{fv} = 1-(P_v-P_f)$ $\qquad P_v > P_f$

1. $\int S_f dP_v$. This is the average value of S_{fv} over all observed conditions for a given forecast only. This value of S_{fv} is called S_f. We can write:

$$\overline{S_f} = \int S_f dP_v$$

The integral, given conditions (i) and (ii), becomes:

$$\overline{S_f} = \int_{P_v=0}^{P_v=P_f} (1-[P_f-P_v])\, dP_v + \int_{P_v=P_f}^{P_v=1} (1-[P_v-P_f])\, dP_v$$

Integration gives:

$$\overline{S_f} = P_f - P_f^2 + 0.5 \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

2. $\int S_v dP_f$. This is the average value of S_{fv} over all forecast conditions for a given observation. This value of S_{fv} is called S_v. A similar equation to equation 3 can be deduced:

$$\overline{S_v} = P_v - P_v^2 + 0.5 \quad \ldots\ldots\ldots\ldots\ldots (4)$$

3. $\iint S_{fv} dP_f dP_v$. This is the grand average of S_{fv} over all forecasts and observations. Noting that $\int S_f dP_v$ is given by equation (3) and $\int S_v dPf_f$ is given by equation (4), then:

$$\int_0^1 \int_0^1 S_{fv} dP_f dP_v = \overline{S_{fv}} = \frac{2}{3} \quad \ldots\ldots\ldots\ldots\ldots (5)$$

This gives from (2) the normalised LEPS1 score:

$$S'_{fv} = 1-(|P_f-P_v|)-1.5(P_f-P_f^2+0.5)(P_v-P_v^2+0.5) \quad \ldots\ldots\ldots\ldots (6)$$

From now onwards the subscript fv will be dropped from S.

2.1 Properties of the LEPS1 S' score

1. The maximum value of the non normalised LEPS1 score, S, (from equation 1b) is 1 when $P_f=P_v$ and its minimum is zero when $P_f=1$ and $P_v=0$, or $P_v=0$ and $P_f=1$. For S', the maximum value from equation 6 is 0.625 and the minimum -0.375. Thus the largest positive score has a value 1.67 times larger than that of the modulus of the most negative score.

2. S' is **doubly equitable**: i.e. it has a chance value of zero for any specific forecast or observation so that the repeated occurrence of a given forecast or observation gives an expected zero score.

3. The S' scores are symmetrical; thus a forecast with a value of the cumulative frequency distribution A that is associated with an observation of value B has the same score as an observation of value A and a forecast of value B. This desirable characteristic is only found with a doubly equitable scoring system.

4. Although S is linear with error in the cumulative probability distribution, the normalised form S' is non-linear (equation 6).

5. The score for a correct forecast of a category near the centre of a distribution is much less than one for a correct forecast near an extreme. For a continuous distribution, when $P_v=P_f=0.1$, S'=0.478 but when $P_v=P_f=0.5$, S'=0.156. This arises because a correct forecast of a value at the boundary of decile 1 and 2 is much further away, in the cumulative probability distribution, from the chance value of the subsequent observation than is a correct forecast of a value placed at the centre of the distribution.

2.2 Forecasts of Continuous variables

For forecasts or simulations of continuous variables, such as those produced by multiple regression methods, equation 6 is applied as it stands. It is only necessary to associate the forecast and observed value with given values of P in their respective cumulative frequency distributions. Here these are assumed to be the same.

2.3 Forecasts of Categorical variables

For forecasts issued in discrete categories, it is best to calculate a mean score averaged over the category. Table 1a (overleaf) shows the LEPS1 quint table, calculated in this way, quoted to two decimal places. This is a square matrix as the forecast and observed categories are defined as identical ranges of the cumulative probability distribution. Categorical forms of LEPS are called LEPSCAT by WF. Table 1b shows Table 1a in the rounded form used in WF; this is rescaled so that the average score for the five correct quint categories is exactly 4, the rescaling used by the FP system in Folland et al. (1986). Table 1c shows the FP quint table for comparison; it is also doubly equitable. Conversion of LEPS1CAT scores in Table 1a to those in Table 1b cannot be achieved precisely; slightly different choices of rounding to one decimal place in Table 1b are possible as it is necessary to make the sums of both rows and columns zero.

2.4 Probability Forecasts

The scores assume that the probability forecast for each category is created from a separate categorical forecast for each category multiplied by a weight equal to the forecast probability (as discussed by WF). So for probability forecasts of five equiprobable categories, if the observation was in quint 4 and the forecast probabilities for quints 1-5 were p_i, $1 \le i \le 5$, the LEPS1PROB score derived from Table 1b is:

$$S_{prob} = -3.4p_1 - 2.3p_2 + 0.1p_3 + 2.9p_4 + 2.7p_5 \quad \ldots\ldots\ldots\ldots \quad (7)$$

Table 1a LEPS1CAT quint table (not rescaled).

		OBSERVATION				
FORECAST		Q1	Q2	Q3	Q4	Q5
	Q1	0.417	0.178	-0.057	-0.222	-0.317
	Q2	0.178	0.184	0.009	-0.149	-0.222
	Q3	-0.057	0.009	0.097	0.009	-0.057
	Q4	-0.222	-0.149	0.009	0.184	0.178
	Q5	-0.317	-0.222	-0.057	0.178	0.417

Table 1b LEPS1CAT quint scores rescaled and rounded as in Ward and Folland (1991).

		OBSERVATION				
FORECAST		Q1	Q2	Q3	Q4	Q5
	Q1	6.4	2.7	-0.8	-3.4	-4.9
	Q2	2.7	2.9	0.1	-2.3	-3.4
	Q3	-0.8	0.1	1.4	0.1	-0.8
	Q4	-3.4	-2.3	0.1	2.9	2.7
	Q5	-4.9	-3.4	-0.8	2.7	6.4

Table 1c Rounded, rescaled FP quint scores as in Folland et al. (1986)

FORECAST	OBSERVATION				
	Q1	Q2	Q3	Q4	Q5
Q1	5.1	1.0	-1.4	-2.3	-2.4
Q2	1.0	3.4	-0.1	-2.0	-2.3
Q3	-1.4	-0.1	3.0	-0.1	-1.4
Q4	-2.3	-2.0	-0.1	3.4	1.0
Q5	-2.4	-2.3	-1.4	1.0	5.1

2.5 Problems with LEPS1 scores

Figure 1a shows, for the continuous case, how S' (not rescaled) varies with P_v for selected values of P_f, $0 \leq P_f \leq 0.5$. For $P_f > 0.5$ (not shown), the curves have a symmetrical behaviour: thus for $P_f=1$, the curve is the mirror image of that for $P_f=0$ with its maximum value at $P_v=1$ and minimum at $P_v=0$. Careful scrutiny of Figure 1a shows that for $P_f=0.3$, 0.4 and 0.5 the minimum score does not coincide with the most erroneous forecast, i.e. the largest distance between P_f and P_v. Thus under these circumstances it is possible for the score to increase, by a small amount, as the error in the forecast increases. This is shown much more clearly in Figure 1b for $P_v>0.9$ and $P_f=0.5$. The behaviour is confined to values of P_v close to 0 and 1. Analysis of equation 6 shows that the minimum scores for $P_f=0.5$ occur at $P_v=0.056$ and 0.944. "Worse" forecasts than these have a slightly less negative value of S'. For all other values of P_f, the minima are closer to $P_v=1$ or $P_v=0$ though all values of P_f in the range $0.211 < P_f < 0.789$ have dual minima that are not coincident with $P_v=0$ and $P_v=1$ as might be hoped. Larger or smaller values of P_f do have minima at $P_v=0$ and $P_v=1$. This undesirable behaviour led to LEPS2.

3. LEPS2

LEPS2 is, numerically, only slightly different from LEPS1 but uses a different normalising factor to create an equitable scoring system that does not suffer from `bending back'. The following method was used because it was found to solve the problem of bending back. It arose from exploratory studies by Potts (1991) and the author of how LEPS1 changed when more general forms of equation 1 were studied.

The derivation of LEPS2 starts from a general form of LEPS with the first term, 1, in equation 1b replaced by the variable a. As a becomes very large, S' becomes independent of a and a new normalisation of S as

defined in equation 1 appears as follows. The general form of LEPS is:

$$S = a - |P_f - P_v| \quad \ldots\ldots\ldots\ldots\ldots (8a)$$

Normalise equation 8a using equations 2a and 2b and integrate in the same way as was done for LEPS1 in Section 2. This gives mean values of S_f, S_v and S_{fv}:

$$\overline{S_f} = a - 0.5 + P_f - P_f^2$$

$$\overline{S_v} = a - 0.5 + P_v - P_v^2$$

$$\overline{S_{fv}} = a - \frac{1}{3}$$

Thus from equation 2b:

$$S' = a - |P_f - P_v| - \frac{(a - 0.5 + P_v - P_v^2)\,(a - 0.5 + P_f - P_f^2)}{a - \frac{1}{3}}$$

Now let $P_v^2 - P_v + 0.5 = \Delta_1$ and $P_f^2 - P_f + 0.5 = \Delta_2 \ldots\ldots\ldots (8b)$.

Assume that a is large. Substituting:

$$S' = a - |P_f - P_v| - \frac{(a - \Delta_1)\,(a - \Delta_2)}{a - \frac{1}{3}}$$

This is approximately

$$S' \approx a - |P_f - P_v| - \left[a + \frac{1}{3}\right]\left[1 + \frac{\Delta_1 \Delta_2}{a^2} - \frac{\Delta_1}{a} - \frac{\Delta_2}{a}\right]$$

$$S' \approx -|P_f - P_v| - \frac{\Delta_1 \Delta_2}{a} + \Delta_1 + \Delta_2 - \frac{1}{3} - \frac{1}{3}\frac{\Delta_1 \Delta_2}{a^2} + \frac{1}{3}\frac{\Delta_1}{a} + \frac{1}{3}\frac{\Delta_2}{a}$$

Let $a \to \infty$ noting that $\Delta = 0(1)$. Add and subtract 1:-

$$S' = 1 - |P_f - P_v| + \Delta_1 + \Delta_2 - 1\tfrac{1}{3} \quad \ldots\ldots\ldots\ldots\ldots\ldots (9a)$$

The first two terms give the non-normalised score S with a=1 as in LEPS1. Now for this form $\overline{S_f}$ is now given by equation 3, $\overline{S_v}$ by equation 4 and $\overline{S_{fv}}$ by equation 5, i.e.:

$$\overline{S_f} = P_f - P_f^2 + 0.5; \quad \overline{S_v} = P_v - P_v^2 + 0.5; \quad \overline{S_{fv}} = \frac{2}{3}$$

From Equation 8b note that

$$\Delta_1 = 1 - \overline{S_f}\ ; \qquad \Delta_2 = 1 - \overline{S_v}$$

Then equation 9a is readily shown to be:

$$S' = 1 - |P_f - P_v| - \overline{S}_f - \overline{S}_v + \overline{S}_{fv} \ldots\ldots\ldots\ldots (9b)$$

$$S' = 1 - |P_f - P_v| - \frac{1}{3} + P_f^2 + P_v^2 - P_f - P_v \ldots\ldots\ldots\ldots\ldots\ldots (9c)$$

Equation 9b gives the same basic score as LEPS1 but with a different, apparently linear, normalisation factor. However normalisation is non-linear in terms of P_f and P_v as shown by equation 9c. The magnitudes of these normalised scores are closely similar to those of LEPS1, as shown below, so that the meaning of the normalised scores is nearly the same as for LEPS1. We call this version LEPS2. It can be regarded conceptually as **almost** a measure of the linear error of the forecast in the cumulative probability distribution relative to a chance value. The similarity of LEPS1 and LEPS2 is illustrated by comparing the non-rescaled LEPS2CAT quint table (Table 2a) with the corresponding LEPS1CAT quint table (Table 1a).

Table 2a LEPS2CAT non-rescaled quint scores

		OBSERVATION				
FORECAST		Q1	Q2	Q3	Q4	Q5
	Q1	0.427	0.174	-0.067	-0.227	-0.307
	Q2	0.174	0.187	0.013	-0.147	-0.227
	Q3	-0.067	0.013	0.107	0.013	-0.067
	Q4	-0.227	-0.147	0.013	0.187	0.174
	Q5	-0.307	-0.227	-0.067	0.174	0.427

Table 2b shows the LEPS2CAT terce table. The terce scores have a simple structure: when calculated to higher resolution, if the scores that equal -0.037 are rescaled to equal -1, all other scores are integers between

-7 and +8. Integral relationships also exist throughout the quint table when the scores equal to 0.013 are rescaled to equal +1.

Table 2b LEPS2CAT non-rescaled terce scores

		OBSERVATION		
FORECAST		T1	T2	T3
	T1	0.296	-0.037	-0.259
	T2	-0.037	0.074	-0.037
	T3	-0.259	-0.037	0.296

3.1 Other Properties of LEPS2

1. Figure 2a shows the continuous form of LEPS2 (LEPS2CONT) in the same format as the LEPS1 scores in Figure 1. Curves for P_f>0.5 (not shown) are again mirror images of those shown, that for P_f=0.7 for instance being the mirror image of that for P_f=0.3. By differentiating equation 9b it is readily shown that the two minima in the scores always occur at P_v=0 and P_v=1 for all P_f, i.e. at these points:

$$\frac{\partial S'}{(\partial P_v)_{P_f}} = 0$$

Thus for a given forecast the worst possible outcome gives asymptotically the worst possible score. This asymptotic property of LEPS2, and the fact that "bending back" is eliminated, is a marked improvement on LEPS1 and a desirable characteristic. Potts (1991) discusses this in much more detail.

2. For a given *state* of P_f relative to P_v, i.e., whether (i) $P_f > P_v$ or (ii) $P_f < P_v$, the curves of S' are parallel. In fact all curves in Figure 2a are different segments of the curve which appears in full for P_f=0. Thus the rate of change of S' for a given state of P_f depends only on P_v and is for the quint table, ignoring rescaling factors:-

(a) For $P_f > P_v$

$$\frac{\partial S'}{\partial P_v} = 2P_v \quad \textit{and} \quad \Delta S' = P_{vA}^2 - P_{vB}^2 \qquad (10a)$$

where A and B are the larger and smaller values of P_v respectively.

(b) For $P_f < P_v$

$$\frac{\partial S'}{\partial P_v} = -2+2P_v \quad \text{and} \quad \Delta S' = (P_{vA}-P_{vB})(P_{vA}+P_{vB}-2) \ldots\ldots\ldots\ldots\ldots (10b)$$

Here $\Delta S'$ is the change in S', for a given P_f and state of P_f, as P_v is changed from P_{vA} to P_{vB}. When $P_f = P_v$, the rate of change of S' with P_v changes discontinuously from a positive to a negative value except for P_v=0 and P_v=1.

3.2 Proposed rescaling of LEPS2

If P_v is set equal to P_f in equation 9c, and equation 9c is integrated from zero to 1 then its mean value is 1/3. If we let S''=3S' then the mean S'' score for correct forecasts in the continuous case is 1 while the maximum S'' score becomes 2 and the minimum -1. It is proposed that S'' should be the preferred, rescaled, form of LEPS2. Figure 2b shows Figure 2a rescaled in this way while tables 2c and 2d (overleaf) show rescaled S''scores for quints and terces rounded to two decimal places so that the sum of rows and columns is exactly zero.

The mean expected S'' score for correct categories varies in an interesting way as the number of categories changes. For two equiprobable categories it is 0.5; for terces (Table 2d with scores calculated to higher resolution) it is 2/3, for quints it is 0.8 and for correct forecasts from the continuous distribution it is 1. Thus the average score for "correct" forecasts rises as the coarseness of the equiprobable categories decreases according to the formula:

$$\overline{S''}_{correct} = 1 - \frac{1}{N} \ldots\ldots\ldots\ldots (11)$$

where N is the number of categories.

Table 2c LEPS2CAT rescaled quint S'' scores

		OBSERVATION				
		Q1	Q2	Q3	Q4	Q5
FORECAST	Q1	1.28	0.52	-0.20	-0.68	-0.92
	Q2	0.52	0.56	0.04	-0.44	-0.68
	Q3	-0.20	0.04	0.32	0.04	-0.20
	Q4	-0.68	-0.44	0.04	0.56	0.52
	Q5	-0.92	-0.68	-0.20	0.52	1.28

Table 2d LEPS2CAT rescaled terce S'' scores

		OBSERVATION		
FORECAST		T1	T2	T3
	T1	0.89	-0.11	-0.78
	T2	-0.11	0.22	-0.11
	T3	-0.78	-0.11	0.89

3.3 Other characteristics of S''

Consider two forecasts corresponding to exactly P_f=0.1 and P_f=0.5 in a continuous distribution. Let $P_f \leq P_v$ in both cases. As seen from Figure 2, the **change** in S'' that occurs when we change P_v, say, from 0.5 to 0.9 is the same for both these forecasts because the *state* of P_f (see section 3.2) is the same in both cases. Thus for P_f=0.1 and P_v=0.5, S''=-0.22 and for the correct forecast P_f=P_v=0.5, S''=0.5, an increase of 0.72. If P_v is changed to 0.9, S'' is -0.94 for P_f=0.1 and P_v=0.9 and -0.22 for P_f=0.5 and P_v=0.9. The increase in the score as P_f approaches P_v is again 0.72.

Consider P_f=0.5 and P_f=0.2. Let P_v initially be 0.1 and finally 0.9. The initial and final states of the two forecasts with respect to P_v are different. The initial scores are -0.22 and 0.65 and the final scores -0.22 and -0.85, reductions of zero and 1.5 in S'' respectively that are very different. This behaviour can be seen from figure 2. Thus, as long as the state of P_f does not change, and for any value of P_f in that state, equal increases in the distance between the forecast and the observation, $|P_f-P_v|$, give equal penalties for S'' scores applied to the continuous distribution. This property does not accurately hold for LEPS1, nor does it hold for LEPS2CAT tables like Tables 2c and 2d.

3.4 Estimating percentage "skill"

It is desirable to have a measure of overall skill in terms of a range of percentage skills from 100% to -100%. The problem is to devise a technique that as far as possible does not lose the equitable character of the LEPS2 scores. A method is developed that succeeds for all practical purposes as long as a sufficiently large ensemble of forecasts is used. To achieve a percentage skill range from 100% to -100%, average percentage skill, SK, is defined for continuous, categorical and probability forecasts as:

$$SK = \frac{\sum 100 S''}{\sum S''_m} \quad \ldots\ldots\ldots\ldots\ldots (12)$$

SK is defined relative to a maximum score S''_m given by the observations

because 100% skill is logically the result of forecasting the same category or value as is subsequently observed. Calculation of S''_m depends on whether the numerator is positive or negative. If positive, $\Sigma S''_m$ is the sum of the maximum possible scores given the observations, i.e. those scores when the forecasts are all correct. If negative, $\Sigma S''_m$ is the sum of the moduli of the worst possible scores given each observation.

For continuous forecasts, S''_m is readily calculated from equation 6 by putting $P_v=P_f$ for positive values of S''. For negative values, for a given P_v, the largest negative score that is needed for use in the denominator is found from the value of P_f that is furthest away from P_v in the cumulative probability distribution. This will be the score corresponding to $P_f=1$ or $P_f=0$. (This would not always be strictly true for LEPS1 because of 'bending back').

3.4.1 Expected biases in SK for single forecasts or observations in a given terce or quint category

For a single forecast, SK is certainly not equitable. Consider the terce S'' table (Table 3a) where the problem is particularly severe. SK involves dividing the score by the maximum possible score given the observation if S'' is positive and by the modulus of the minimum possible score given the observation if S'' is negative. Table 3a shows the resulting table for both a single forecast and a single observation:

Table 3a LEPS2 SK percentage skill terce table for a single forecast and a single observation

		OBSERVATION			
FORECAST		T1	T2	T3	mean
	T1	100	-100	-100	-33.3
	T2	-14.1	100	-14.1	23.9
	T3	-100	-100	100	-33.3
	mean	-4.7	-33.3	-4.7	-14.2

The expected mean skill, using equation 12, for the whole table is -14.24%. SK=-33.3% for a single forecast of terce 1 or terce 3 and 23.9% for a single forecast of terce 2. For a single observation of terce 2, the expected value of SK=-33.3% and for single observations of terce 1 or terce 3, the expected value of SK=-4.7%. Table 3b gives SK values for single forecasts and observations of quints and the expected mean values for each forecast and observed quint. Expected SK values are: quints 1 and 5, -11.00%, quints 4 and 2, +7.06 % and quint 3,-15.0%. The average bias for one forecast when all forecast and observed quints are equally

likely is -4.58%, less than that for terces (-14.24%).

Table 3b LEPS2 SK percentage skill quint table for a single forecast and a single observation

		OBSERVATIONS					
		Q1	Q2	Q3	Q4	Q5	mean
FORECAST	Q1	100	92.9	-100	-100	-100	-21.4
	Q2	40.6	100	12.5	-64.7	-73.9	2.9
	Q3	-21.7	7.1	100	7.1	-21.7	14.2
	Q4	-73.9	-64.7	12.5	100	40.6	2.9
	Q5	-100	-100	-100	92.9	100	-21.4
	mean	-11.0	7.1	-15.0	7.1	-11.0	-4.58

3.4.2 Expected biases in SK for pairs of forecasts or observations in a given terce category

Consider the terce SK values in Table 3a. The expected value of SK aggregated over a **pair** of forecasts becomes less positive or negative so that the bias reduces. To demonstrate this, consider pairs of **forecasts** of terce 2. There are 9 possible pairs of **observations**, some repeated, each with equal probability: (i) observed terce 1 and terce 1 (ii) terce 1 and terce 2 (iii) terce 1 and terce 3 (iv) terce 2 and terce 1 (v) terce 2 and terce 2 (vi) terce 2 and terce 3 (vii) terce 3 and terce 1 (viii) terce 3 and terce 2 (ix) terce 3 and terce 3. From Table 2d, the corresponding S'' sums of the first three pairs are (i)-0.22, (ii) 0.11 and (iii) -0.22. Corresponding denominators of equation 12 are (i) a score equal to the sum of the moduli of the scores for two maximally incorrect forecasts of terce 2, 1.56 (because the numerator is negative), (ii) a score equal to the sum of the maximum possible scores when (a) terce 1 was observed (0.89) and (b) when terce 2 was observed (0.22) giving 1.11 (iii) gives the same numbers as (i), i.e. a total of 1.56. The corresponding SK values are -14.10%, 9.91% and -14.10%. Table 4a (overleaf) shows the complete calculation for pairs of forecasts of terce 2 while Table 4b shows the corresponding calculation for pairs of observations of terce 2, allowing for all possible combinations of forecasts. The expected bias in SK for pairs of forecasts of terce 2 is 9.25%, markedly less than half of that for single forecasts of terce 2 (23.93%). For pairs of observations of terce 2, the bias of -22.22% is two thirds of that for single observations of terce 2 (-33.33%). The

initial bias in SK for other categories declines in the same way.

Table 4a LEPS2 SK for pairs of forecasts of terce 2

Combinations of Terces observed	numerator	denominator	SK
1 and 1	-0.22	1.56	-14.10
1 and 2	0.11	1.11	9.91
1 and 3	-0.22	1.56	-14.10
2 and 1	0.11	1.11	9.91
2 and 2	0.44	0.44	100.00
2 and 3	0.11	1.11	9.91
3 and 1	-0.22	1.56	-14.10
3 and 2	0.11	1.11	9.91
3 and 3	-0.22	1.56	-14.10
			mean SK=9.25%

Table 4b LEPS2 SK for pairs of observations of terce 2

Combinations of Terces forecast	numerator	denominator	SK
1 and 1	-0.22	0.22	-100
1 and 2	0.11	0.44	25
1 and 3	-0.22	0.22	-100
2 and 1	0.11	0.44	25
2 and 2	0.44	0.44	100
2 and 3	0.11	0.44	25
3 and 1	-0.22	0.22	-100
3 and 2	0.11	0.44	25
3 and 3	-0.22	0.22	-100
			mean SK=-22.22%

3.4.3 Expected biases for many aggregated forecasts

To illustrate the behaviour of SK as the number of forecasts increases further, Figures 3a and 3b show SK biases for randomly made forecasts of terces and quints for persistently **forecast** individual categories using equation 12. The diagrams are plotted as a function of the number of (independent) forecasts between 1 and 200 using a Monte Carlo sampling technique. Thus for a persistent forecast of terce 3, observations were randomly made of terces 1,2 and 3. Sampling was in fact done for up to 400 aggregated forecasts. Each of the 400 simulations for a given forecast terce or quint was repeated 8000 times, as the standard deviation of SK values is very high compared to their means.

The simulations give SK biases for one forecast close to the theoretical value in each case. For (a) persistent forecasts of terce 1 or 3, and (b) persistent forecasts of terce 2 the following biases in SK are obtained for randomly observed terces: 5 aggregated forecasts: -4.78% and 2.32% (much less than their values -33.33% and 9.25% for one forecast); by 25 forecasts these biases reduce to -1.79% and 0.16%; by 100 forecasts to -0.73% and effectively zero, and by 400 forecasts are very small at -0.47% and zero. For (c) persistent forecasts of quint 1 or 5 (d) quint 2 or 4 (e) quint 3 the biases for randomly observed quints are: 5 aggregated forecasts: -4.54%, 3.35% and -5.87%, 25 aggregated forecasts: -2.22%, 1.42% and -2.78%, 100 aggregated forecasts: -1.25%, 0.67% and -1.43%, 400 aggregated forecasts: -0.55%, 0.33% and -0.72% respectively. Thus SK tends quickly then slowly to zero as the number of aggregated forecasts rises. Similar results apply to the other individual terce observation and forecast categories as the number of independent forecasts increases.

Figure 3c gives the SK values when all forecast and observed terces, and all forecast and observed quints, are randomly sampled. 6400 samples were used. For one forecast, the SK score of -13.59% for terces is close to the theoretical value of -14.24% and the score of -4.59% for quints is close to the theoretical value of -4.58%. For five aggregated forecasts, the biases in SK for (i) randomly forecast and observed terces and (ii) randomly forecast and observed quints are calculated to be -4.03% and -2.89% , for 25 forecasts -1.76% and -1.00%, for 100 forecasts -0.77% and -0.51% and for 400 aggregated forecasts -0.38% and -0.26% respectively.

Figure 3d gives corresponding results, using 6400 samples, for a continuous distribution. The SK bias for one forecast is estimated to be -7.87%, for 5 forecasts -3.39%, for 25 forecasts -1.57%, for 100 forecasts -0.77% and for 400 forecasts -0.43%.

There are some sampling variations in these values; thus biases for all terces and quints in Figure 3c are not exactly consistent with those for individual observed terces or quints shown in Figures 3a and 3b. Similar calculations for persistently observed quints or terces give the same overall picture. However, the detailed behaviour of individual persistently observed quints and terces is not the same as that for persistently forecast quints and terces for small numbers of aggregated forecasts. This can be seen (for one forecast or observation) from Tables 3a and 3b. For about 20 or more forecasts, biases in SK can usually be neglected for terces and even for five forecasts, biases may often not be serious for quints or for forecasts made from a continuous distribution. A broadly similar structure of SK biases is likely for the FP skill scoring system. However, SK scores are not suitable for a single independent forecast as they are insufficiently equitable and the overall negative bias is rather large.

Potts (1991) shows that if the same probability distribution is used for both observations and forecasts but a set of forecasts is issued with a moderately higher variance than that of the observations (the opposite case is more likely) then S' or S'' is slightly increased. Thus LEPS2 could be "played" in this way to slightly increase average scores. This can be overcome by recalculating the cumulative probability distribution of the forecasts to reflect its larger variance compared to the observations. Standardisation of the forecast and observed values by

their respective standard deviations would have the same effect. Potts (1991) contains details of this and other characteristics of LEPS2.

3.5 A second type of skill score

Equation 11 showed that the mean LEPS S'' score for correct forecasts increase from 0.5 when two categories were used to 1 for a continuous distribution. It might be argued that a complete scoring system should allow for the number of categories; there is less information in a forecast of one of two possible categories than in one of the forecast of a specific value taken from a continuous distribution. Whether a mere doubling of the expected scores for a correct forecast is sufficient when the number of categories changes from two to the continuous case (N is infinite) is perhaps unclear. Nevertheless, S'' scores can be converted into SK scores that allow for the number of categories by multiplying the SK scores when positive by 1-1/N as can be seen from equations 11 and 12, where N is the number of categories (N>=2). This would give a maximum possible SK score for a (sufficiently large) perfect set of two category forecasts of 50%, for terces 66.67% and for quints 80%.

3.6 What is a good LEPS2 SK score in long-range forecasting?

There is no definite answer to this question but some preliminary guidance can be given. A forecast of quint 1 and an observation of quint 2, or a forecast of quint 2 and an observation of quint 1 can be regarded as "good" forecast for observations falling on average in the centre of these categories. The SK value for this pair of possibilities is, from Table 3b, 56.5%. So a minimum SK value of at least 50% may be needed, averaged over a set of at least five independent (or equivalently independent) quint forecasts to minimise bias in the estimate, before such a set can be regarded as distinctly "good". This will be harder to achieve for probability forecasts where the appropriate average score may be distinctly less, and a little easier to achieve with terces. It has also been found that "marginally useful" UK long-range forecasts required a 4 year average FP skill score of about 15% for terce or quint categories. It can therefore be surmised that 15% is near the minimum long term LEPS2 SK value of practical value; for seasonal decile forecasts, such as used in Australia (Beard, personal communication), the corresponding SK value could be a little less because the number of categories is greater. A long term FP skill score of 40% was reached in some recent years for UK district temperature quint forecasts (Legg, personal communication); the forecasts were definately of value to a variety of specialised users. Some recent reduction in skill has still not caused a major reduction in their apparent value. So a long term LEPS2 SK value of at least 30% for quints might be regarded as "moderately useful" in long-range forecasting, with a somewhat lower limit for deciles.

4. Conclusions

LEPS2 appears to be a distinct improvement on LEPS1. Its only obvious disadvantage appears to be that its interpretation is less clear than for LEPS1, though the differences in the scores are small. Thus LEPS2 scores *almost* measure the error of the forecast in the cumulative probability distribution relative to a chance level. It is hoped to use LEPS2 for

assessing seasonal tropical long-range forecasts and as one of several tools for quantifying the skill of climate model simulations of atmospheric patterns over the last century. The SK percentage skill score version is considered the most useful form of LEPS2 because it is particularly easy to understand. However SK scores are biased on average negatively; the bias quickly, then much more slowly, decreases towards zero as the number of forecasts increases. SK scores should only be used when about five or more independent forecasts are aggregated, though preferably more than 20 forecasts should be aggregated when terces are used. For sets of grid point meteorological fields the continuous form of LEPS2 should be used. The effective number of independent point forecasts or simulations aggregated in space and time rapidly becomes large for such fields. SK biases will often not be serious for a sufficiently extensive single field as shown by Figure 3d and should become negligible when more than a few fields are assessed together.

Acknowledgements

The author is grateful to Dr J M Potts, recently of the University of Kent, for access to her Phd thesis which contributed considerably to the development of LEPS2. Valuable critical comments from Dr Potts, Mr D Parker and Professor I Jolliffe were received. An unpublished paper by Tony Barnston of the Climate Analysis Center, Washington, DC, also stimulated further thinking.

REFERENCES

Daan, H., 1984: Scoring rules in forecast verification. Short and Medium Range Weather Prediction Research Publication Series No 4, WMO, 62pp.

Folland, C.K., A. Woodcock and L. Varah, 1986: Experimental monthly long-range forecasts for the United Kingdom: Part III, Skill of the monthly forecasts. Met. Mag., 115, 377-395.

Gandin, L.S., and A.H. Murphy, 1992: Equitable skill scores for categorical forecasts. Mon. Weath. Rev., 120, 361-370.

Potts, J.M., 1991: Statistical methods for the comparison of spatial patterns in meteorological variables. Phd Thesis, Univ. of Kent.

Ward, M.N., and C.K. Folland, 1991: Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperature. Int. J. Climatol., 11, 711-743.

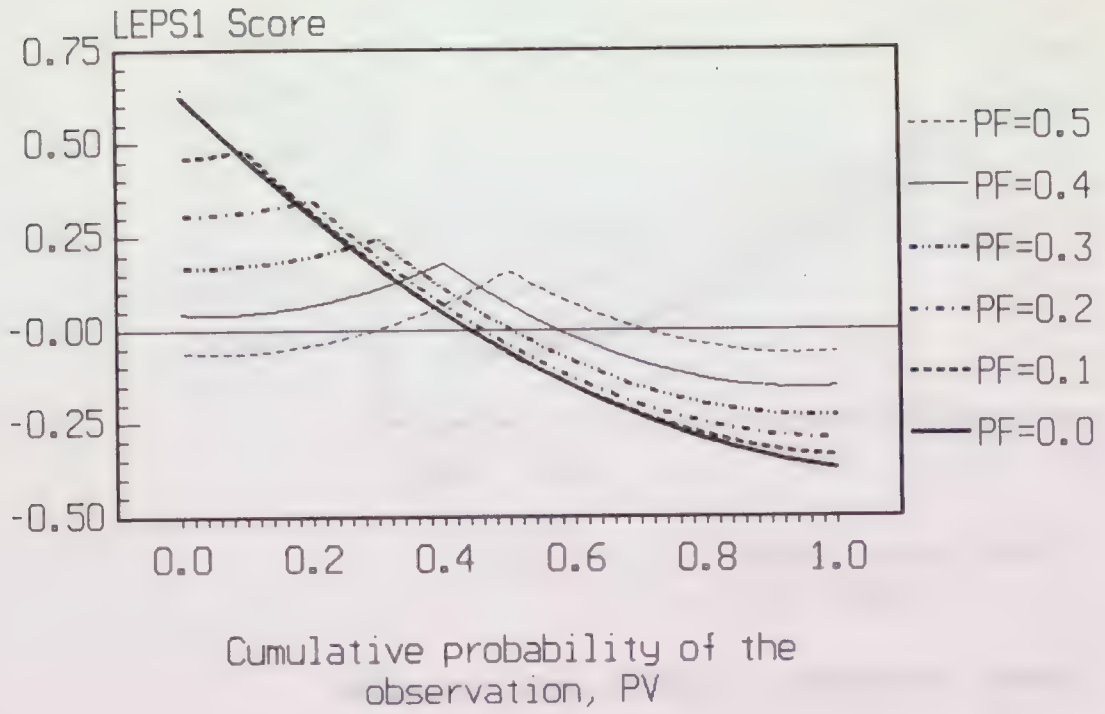

Figure 1a **LEPS1 continuous scores, not rescaled.**

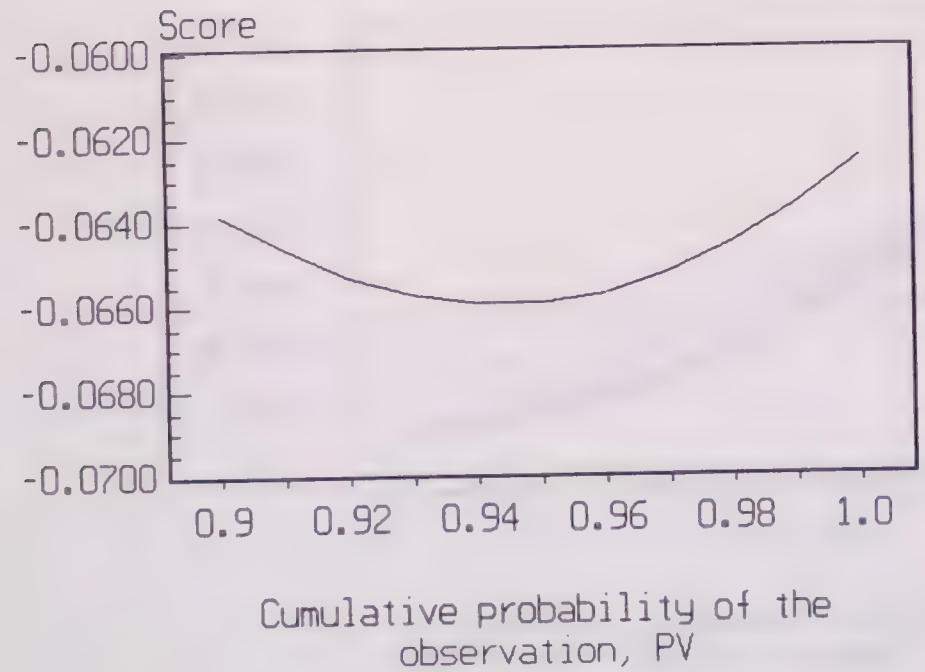

Figure 1b **Figure 1a expanded for P_v>=0.9 and P_v=0.5**

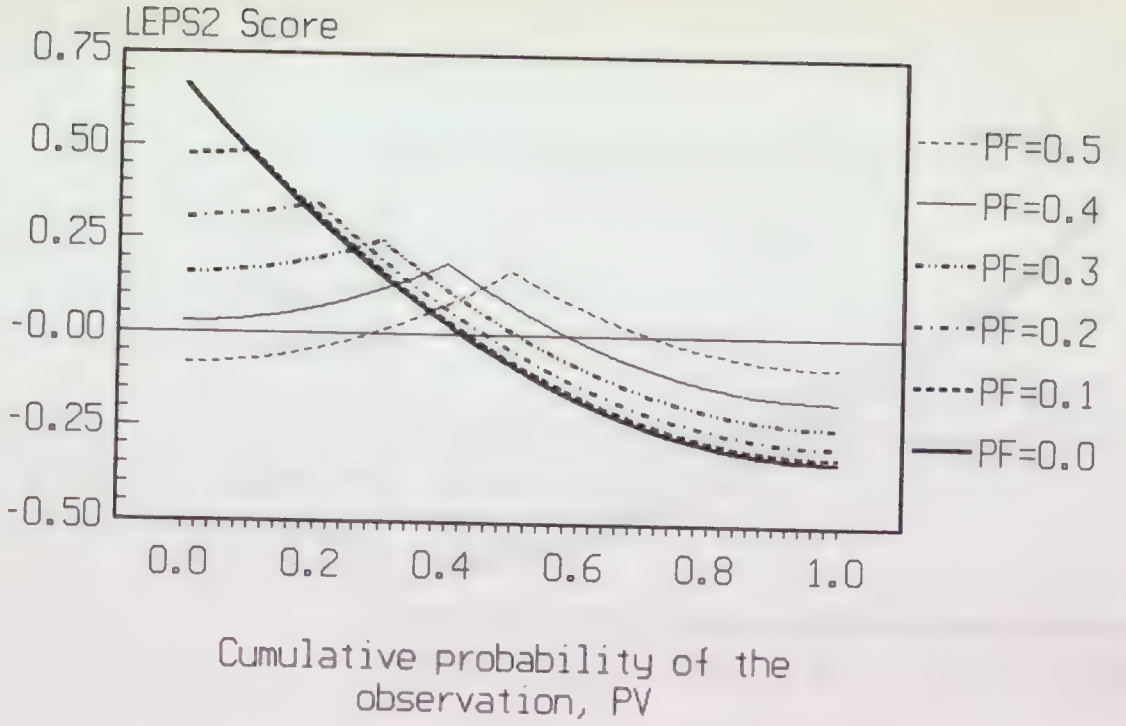

Figure 2a **LEPS2 continuous scores, not rescaled.**

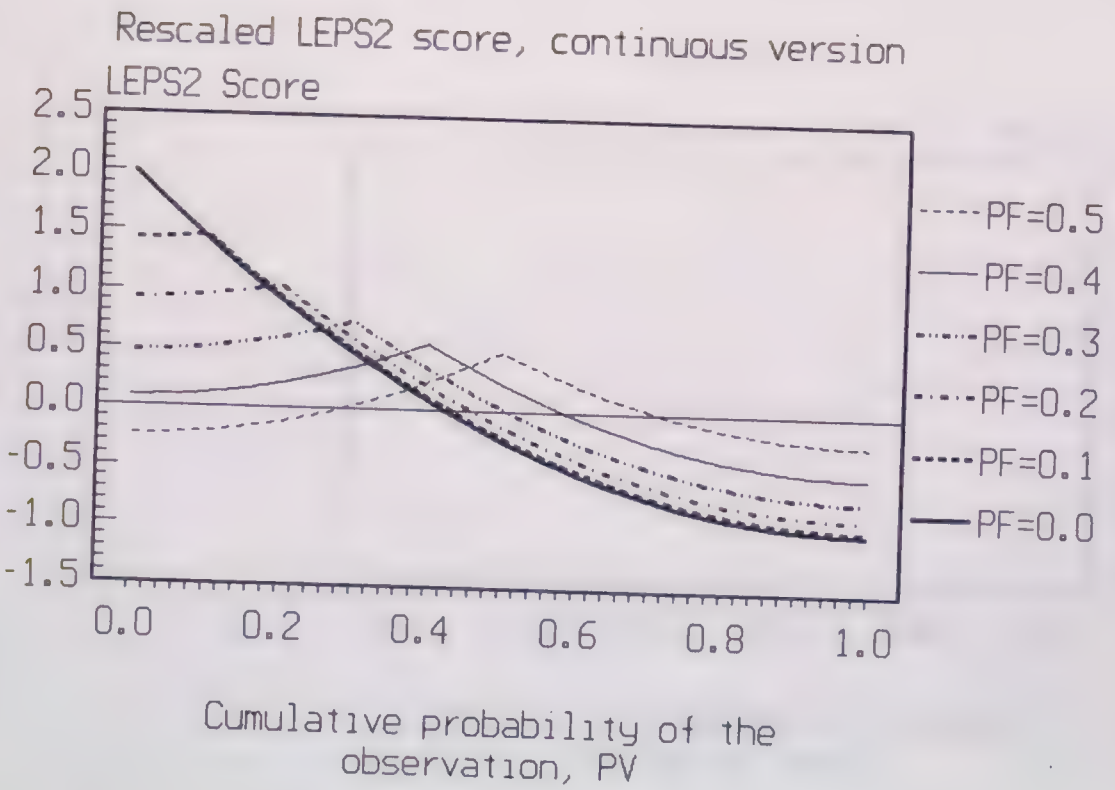

Figure 2b **LEPS2 continuous scores, rescaled**

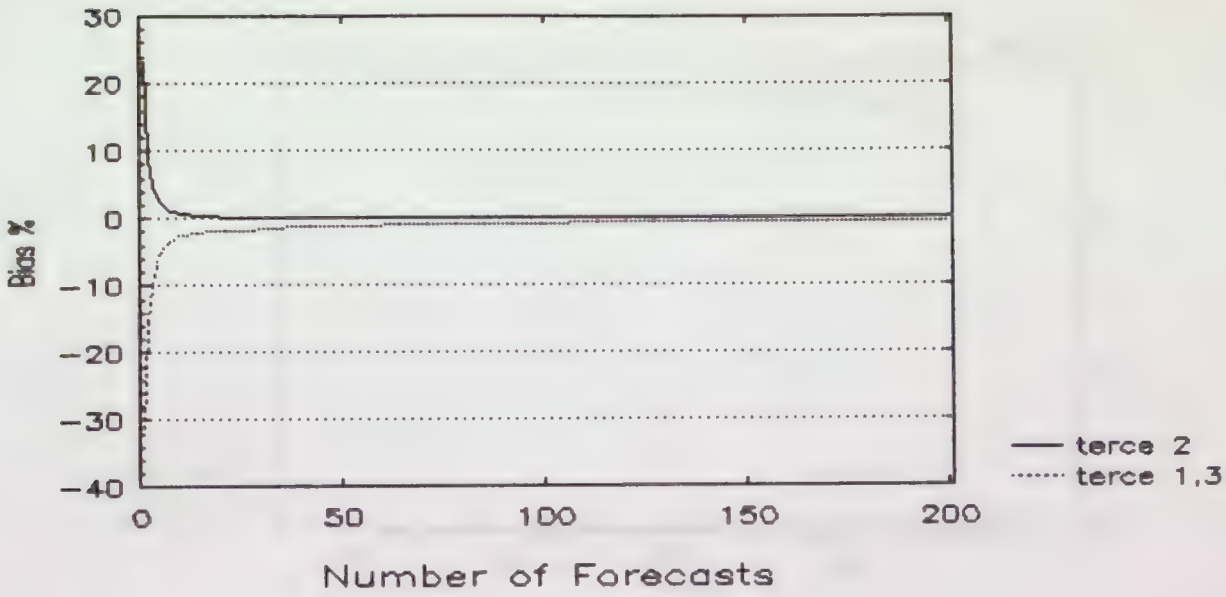

Figure 3a Biases in the expected percentage SK skill score for individual persistently forecast terces and randomly observed terces.

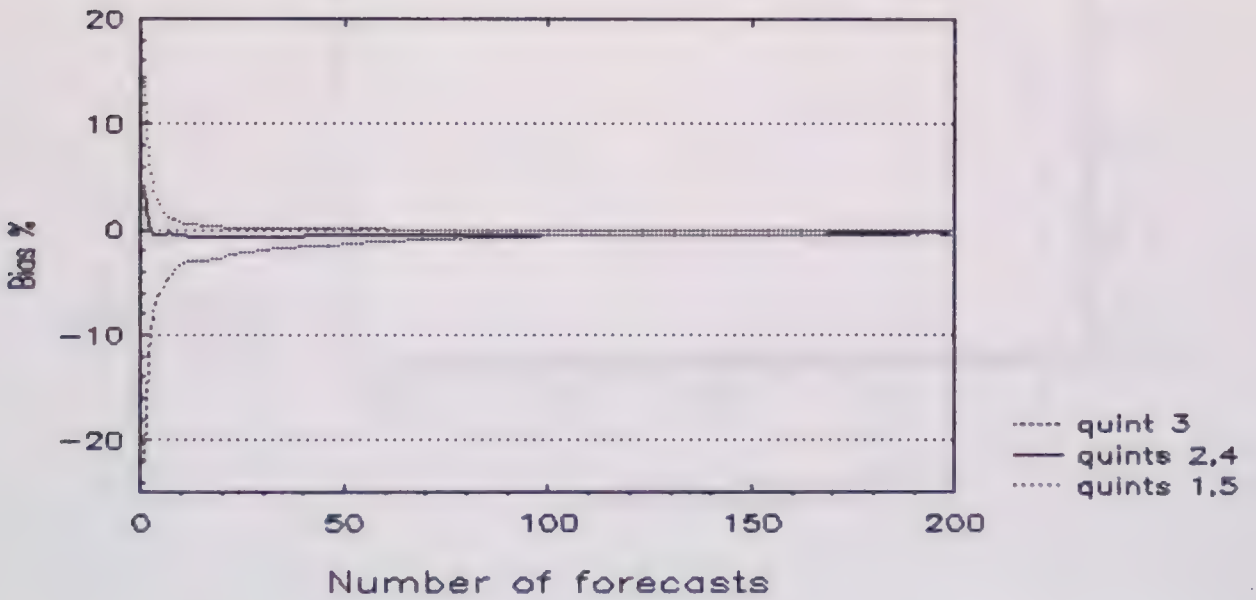

Figure 3b As Fig. 3a except for individual persistently forecast quints and randomly observed quints.

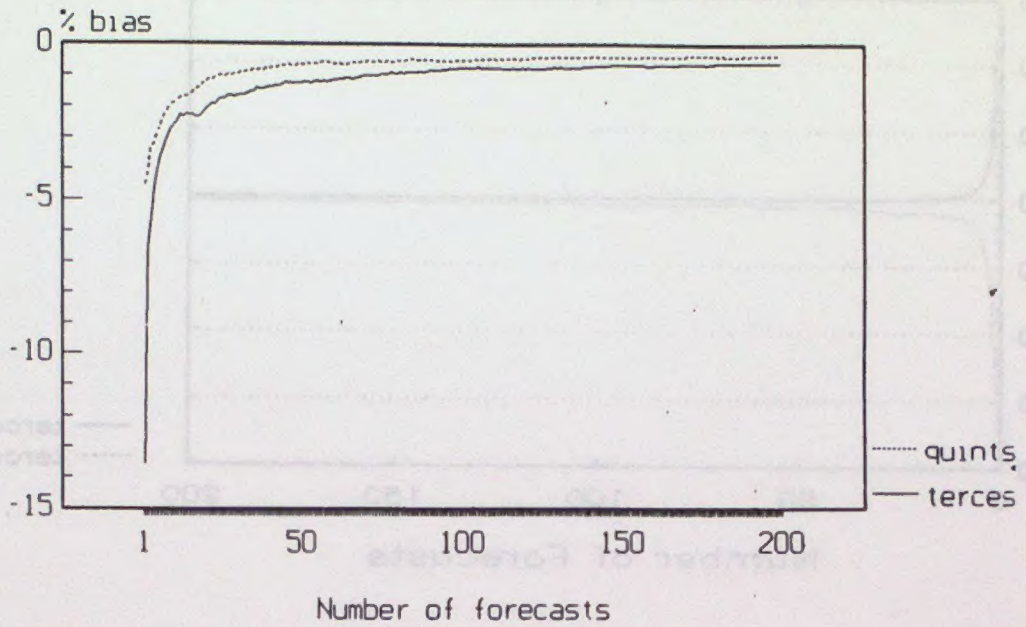

Figure 3c As Fig. 3a but for all terces and quints both randomly observed and forecast.

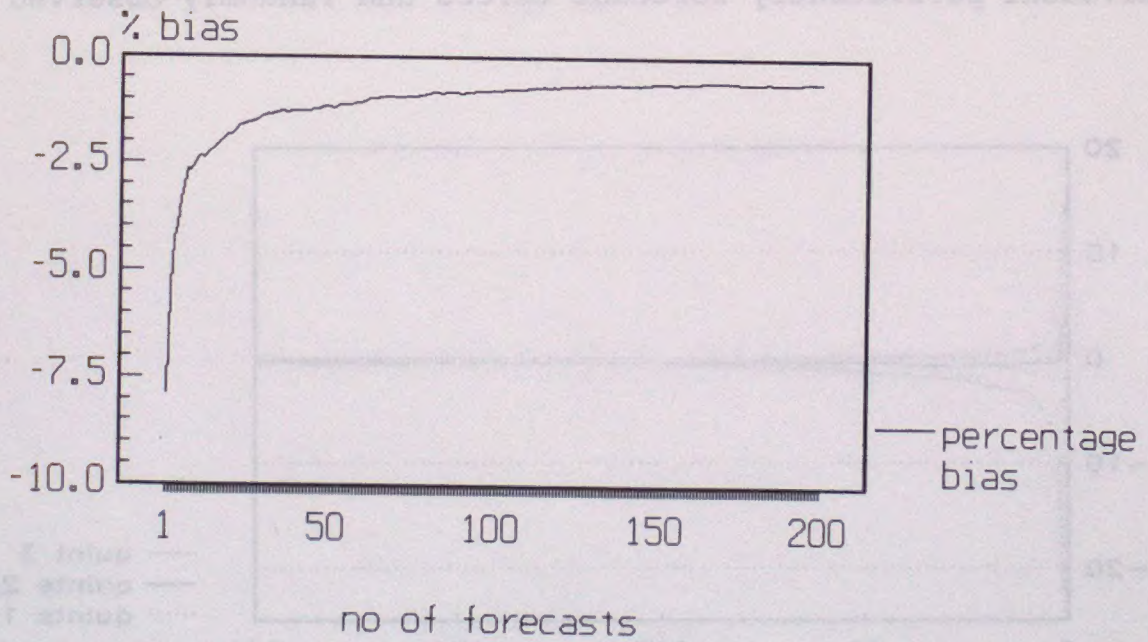

Figure 3d As Fig. 3a but for a continuous distribution of forecasts and observations.

CLIMATE RESEARCH TECHNICAL NOTES

CRTN 1	Oct 1990	Estimates of the sensitivity of climate to vegetation changes using the Penman-Monteith equation. P R Rowntree
CRTN 2	Oct 1990	An ocean general circulation model of the Indian Ocean for hindcasting studies. D J Carrington
CRTN 3	Oct 1990	Simulation of the tropical diurnal cycle in a climate model. D P Rowell
CRTN 4	Oct 1990	Low frequency variability of the oceans. C K Folland, A Colman, D E Parker and A Bevan
CRTN 5	Dec 1990	A comparison of 11-level General Circulation Model Simulations with observations in the East Sahel. K Maskell
CRTN 6	Dec 1990	Climate Change Prediction. J F B Mitchell and Qing-cun Zeng
CRTN 7	Jan 1991	Deforestation of Amazonia - modelling the effects of albedo change. M F Mylne and P R Rowntree
CRTN 8	Jan 1991	The role of observations in climate prediction and research. D J Carson
CRTN 9	Mar 1991	The greenhouse effect and its likely consequences for climate change. D J Carson
CRTN 10	Apr 1991	Use of wind stresses from operational N.W.P. models to force an O.G.C.M. of the Indian Ocean. D J Carrington
CRTN 11	Jun 1991	A new daily Central England Temperature series, 1772-1991. D E Parker, T P Legg and C K Folland
CRTN 12	Jul 1991	Causes and predictability of Sahel rainfall variability. D P Rowell, C K Folland, K Maskell, J A Owen, M N Ward
CRTN 13	Jul 1991	Modelling changes in climate due to enhanced CO_2, the role of atmospheric dynamics, cloud and moisture. C A Senior, J F B Mitchell, H Le Treut and Z-X Li

CLIMATE RESEARCH TECHNICAL NOTES

CRTN 14	Sep 1991	Sea temperature bucket models used to correct historical SST data in the Meteorological Office. C K Folland
CRTN 15	Aug 1991	Modelling climate change, and some potential effects on agriculture in the U.K. P R Rowntree, B A Callander and J Cochrane
CRTN 16	Aug 1991	The Boreal Forests and Climate G Thomas and P R Rowntree
CRTN 17	Aug 1991	Development of a Stratosphere-Troposphere Data Assimilation System. R Swinbank.
CRTN 18	Sep 1991	A study of asynchronous coupling using a simple climate model. M K Davey.
CRTN 19	Sep 1991	The Oceanic Carbon Cycle. N K Taylor
CRTN 20	Nov 1991	Worldwide ocean-atmosphere surface fields in Sahel wet and dry years using provisionally corrected surface wind data. M N Ward
CRTN 21	Dec 1991	Coupled tropical ocean global atmosphere models at the UKMO. M K Davey, C Gordon, S Ineson and S Lawrence
CRTN 22	Dec 1991	Empirical parameterisation of tropical ocean-atmosphere coupling. M Allen, M K Davey, D L T Anderson and P D Killworth
CRTN 23	Feb 1992	The temporal evolution of Equatorial Currents in the Indian Ocean D L T Anderson, D J Carrington, R A Corry and C Gordon
CRTN 24	Feb 1992	Stratospheric analyses provided by the U.K. Meteorological Office M J Bailey, A O'Neill and V D Pope
CRTN 25	Feb 1992	Modelling interannual variability in the Indian Ocean using momentum fluxes from the UKMO and ECMWF operational weather analyses. D L T Anderson and D J Carrington

CLIMATE RESEARCH TECHNICAL NOTES

CRTN 26	Mar 1992	A GCM simulation of the impact of Amazonian deforestation on climate using an improved canopy representation J Lean and P R Rowntree
CRTN 27	Mar 1992	The parameterization of rainfall interception in GCMs A J Dolman and D Gregory
CRTN 28	Jun 1992	Development of worldwide marine data eigenvectors since 1985 A W Colman
CRTN 29	Jun 1992	A tropical ocean model with reduced physics M K Davey
CRTN 30	Jun 1992	International Temperature Workshop, Boulder, Colorado, USA, 16 January 1992 Edited by D E Parker
CRTN 31	Jul 1992	Simulation of clear-sky outgoing longwave radiation over the oceans using operational analyses A Slingo and M J Webb
CRTN 32	Sep 1992	A prediction of the transient response of climate J M Murphy
CRTN 33	Nov 1992	LEPS scores for assessing climate model simulations and long-range forecasts. C K Folland